[1]Simulation of Incorporating the Variance from Missing Data in Census Coverage Measurement Variance Estimates

Richard A. Griffin
U.S. Census Bureau

Abstract

For the Census 2000 Accuracy and Coverage Evaluation (A.C.E.), variance due to missing data imputation was incorporated into the jackknife variance estimates by the re-calculation of the missing data adjustments for missing P-sample match status, E-sample correctness of enumeration, and residence probability for each jackknife replicate. These missing data adjustments were done by simple cell methodology whereby, for example, for each cell the weighted proportion of P-sample matches for resolved cases was used as the imputed probability of a match for unresolved cases in that cell. For the 2010 Census
Coverage Measurement (CCM) Survey, imputation will be done using logistic regression. A possibility for variance estimation is re-computing the estimated logistic regression coefficients for each jackknife replicate to account for missing data variance. This paper documents a simplified empirical study using simulated data and the R programming language to evaluate the bias and variance properties of a jackknife variance estimate that attempts to include variance due to a missing data imputation using logistic regression.

Key Words: Logistic Regression Imputation Variance Estimation

1. Background

For the Census 2000 Accuracy and Coverage Evaluation (A.C.E.), variance due to missing data imputation was incorporated into the jackknife variance estimates by the re-calculation of the missing data adjustments for missing P-sample match status, E-sample correctness of enumeration, and residence probability for each jackknife replicate. These missing data adjustments were done by simple cell adjustment whereby, for example, for each cell the weighted proportion of P-sample matches for resolved cases was used as the imputed probability of a match for unresolved cases in that cell. Lee (2002) discusses the theory for this approach and provides details and results from an empirical study implemented on two artificial populations. The imputation methods discussed are ratio imputation (cell imputation is a special case) and nearest neighbor hot deck. For the 2010 Census Coverage Measurement (CCM) Survey, imputation will be done using logistic regression. While it makes intuitive sense that re-computing the estimated logistic regression coefficients for each jackknife replicate may properly account for missing data variance, we were not able to find documentation of theory or an empirical study specifically dealing with missing data variance estimation for the case of logistic

[1] *This report is released to inform interested parties of research and to encourage discussion. Any views expressed on statistical, methodological, technical, or operational issues are those of the author and not necessarily those of the U.S. Census Bureau.*

regression based imputation. This paper documents a simplified empirical study using simulated data and the R programming language to evaluate the bias and variance properties of a jackknife variance estimate that attempts to include variance due to missing data imputation.

Section 2 gives details of the simulation and section 3 provides results. Section 4 is a summary and section 5 provides references. The R programs along with discussion of the reason for sections of code used for this empirical study are available from the author.

2. Empirical Study Using Simulated Data

Consider a population of 1000 persons each with an independent probability of capture in the Census. Define this probability as follows:

$$p_j = \frac{\exp(.5 + .8X_j)}{1 + \exp(.5 + .8X_j)}, \quad (1)$$

where X~N(0,1). This is one of the models used to generate capture probabilities in Alho (1990).

N = 1000 independent observations of X from N (0,1) are generated. Using an independent random draw from the Uniform (0,1) distribution and p_j, each person is classified as either 1 (captured in the Census) or 0 (not captured in the Census). The P-sample is then defined as an independent simple random sample without replacement of size 100 from N.

The logistic regression model used for estimation is as follows:

Let y_j be a random dependent variable defined as 1 if P-sample person j is enumerated in the Census (i.e., matches) and 0 otherwise.

π_j = the probability that $y_j = 1$. Note that π_j is not the same as p_j, which is only used in setting up the simulation population classifying each person as either 1 (captured in the Census) or 0 (not captured in the Census) as well as determining response or no response (see section 2.2) for census capture. p_j is not available for estimation.

$$\ln\left(\frac{\pi_j}{1-\pi_j}\right) = \beta_0 + \beta_1 X_j + \varepsilon_j \quad (2)$$

and once the model is fit, $\hat{\pi}_j = \frac{\exp(\hat{\beta}_0 + \hat{\beta}_1 X_j)}{1 + \exp(\hat{\beta}_0 + \hat{\beta}_1 X_j)}$

The dual system estimator used is as follows:

$$\hat{N} = \sum_{j \in Census} \frac{1}{\hat{\pi}_j} \quad (3)$$

Note that for this study we are assuming all census enumerations are data-defined and correct enumerations. Also, all results are conditional on the population of size N created as described.

2.1 Simulation 1 – Complete Response

The first simulation is for jackknife variance estimation if all P-sample persons are resolved as either captured (i.e., matched) or not captured in the Census. One thousand simulations of a P-sample of 100 persons from the universe of 1000 persons are done. For each sample, logistic regression model (2) is fit and $\hat{N}$ is computed as in equation (3) and stored. Then for each of 100 jackknife replicates, c = 1,…,100, a P-sample person is sequentially removed leaving 99 persons for whom logistic regression model (2) is again fit. The estimated coefficients from this fitting are stored for each replicate. Using these estimated coefficients $\hat{N}_{(c)}$ is computed leaving out P-sample person c. The jackknife variance estimate is computed as follows;

$$\hat{V}_{jack}(\hat{N}) = .9(\frac{99}{100})\sum_{c=1}^{100}(\hat{N}_{(c)} - \sum_{c=1}^{100}\frac{\hat{N}_{(c)}}{100})^2 \quad (4)$$

2.2 Simulation 2 – Some Nonresponse, Impute Enumerated or not Enumerated in the Census

For each of the N population records, a response probability, equal to $(.96)p_j + (.8)(1-p_j)$ (see equation (1)), is calculated and using the independent random draw from the Uniform (0,1) distribution and this response probability, each person is classified as either 1 (a respondent) or 0 (a nonrespondent). This provides approximately the amount of nonresponse to match status observed in Census 2000 A.C.E.

The CCM Survey includes a listing of housing units in sample blocks and matching the housing units to the Census list of housing units. Define a new variable Z defined as 1 if the housing unit a person lives in matched to the Census and 0 otherwise. Z is known to be highly correlated with a P-sample person matching or not to the Census. Z is only available for persons in the P-sample so that it can only be used for imputation (not available for all census enumerated persons so cannot be used in equation (3) for estimation). For our simulation population, this variable is set to 0 for persons with census capture probability, p_j, less than or equal to 0.5 (about 25% of the simulated universe) and set to 1 otherwise.

The second simulation allows for missing data and uses logistic regression for missing data imputation as well as for estimation after imputation.

The logistic regression model used for imputation is as follows:

Let y_j be a random dependent variable defined as 1 if P-sample person j is enumerated in the Census (i.e., matches) and 0 otherwise.

π_j = the probability that $y_j = 1$.

$$\ln(\frac{\pi_j}{1-\pi_j}) = \beta_0 + \beta_1 X_j + \beta_2 Z_j + \varepsilon_j \quad (5)$$

and once the model is fit, $\hat{\pi}_j = \dfrac{\exp(\hat{\beta}_0 + \hat{\beta}_1 X_j + \hat{\beta}_2 Z_j)}{1 + \exp(\hat{\beta}_0 + \hat{\beta}_1 X_j + \hat{\beta}_j Z_j)}$ (6)

One thousand simulations of a P-sample of 100 persons are done.
For each sample, logistic regression model (5) is fit using the respondents in that sample. Then equation (6) is used to estimate the match probability for each nonrespondent person. Using the independent random draw from the Uniform (0,1) distribution and this match probability, each nonrespondent person is classified as either 1 (a match) or 0 (a nonmatch).

Using these imputed values for the nonrespondents and the observed values for the respondents, logistic regression model (2) is fit and $\hat{N}$ is computed as in equation (3) and stored.

Next for each of 100 jackknife replicates, c = 1,…,100, a P-sample person is sequentially removed leaving 99 persons. If the removed person was a respondent, logistic regression model (5) is fit using the remaining respondents and used for imputation for the nonrespondents. Logistic regression model (2) is then fit using these imputed values for the nonrespondents (or the imputed values already obtained for estimation if the removed person was a nonrespondnet) and the observed values for the respondents for the 99 persons in that jackknife replicate. It is not necessary to impute again if the removed person was a nonrespondent since in that case, the remaining respondents are the same as for estimation. The estimated coefficients from this fitting are stored for each replicate. Using these estimated coefficients $\hat{N}_{(c)}$ is computed. The jackknife variance estimate is computed using equation (4).

For comparison purposes this simulation is run again using a jackknife variance estimator that treats imputed values the same as observed values (no re-imputation for each jackknife replicate). Once the imputation is done for a sample, treat the imputed values as observed and then apply equation (4) as in simulation 1.

2.3 Simulation 3 – Some Nonresponse, Impute by Creating Two Records with Weights

The third simulation also allows for missing data and uses logistic regression for missing data imputation as well as for estimation after imputation. One thousand simulations of a P-sample of 100 persons are done.

For each sample, logistic regression model (5) is fit using the respondents in that sample. Then equation (6) is used to estimate the match probability for each nonrespondent person. Instead of imputing either 1 (a match) or 0 (a nonmatch) using the random draw from the Uniform (0,1) distribution as in simulation 2, two records are created for each nonrespondent. Given an imputed match probability of $\hat{\pi}_j$, one record has a dependent response variable of 1 (matched) and a weight of $\hat{\pi}_j$ and the second record has a dependent response variable of 0 (nonmatch) and a weight of $1-\hat{\pi}_j$. Logistic regression model (2) is fit using these weights for the records representing the nonrespondents and a weight of 1 for the records representing respondents. $\hat{N}$ is computed as in equation (2) using weighted logistic regression and stored.

Next for each of 100 jackknife replicates, c = 1,…,100, a P-sample person is sequentially removed leaving 99 persons. For each jackknife replicate of 99 persons, the entire process of fitting logistic regression model (5) using the respondents, using the fitted model to create two records for nonrespondents and then fitting logistic regression model (2) to compute $\hat{N}_{(c)}$ is repeated for each jackknife replicate. Note that if the removed person was a nonrespondent it was not necessary to repeat the entire process since the remaining respondents were the same as from the full sample. Here the process was repeated even if the removed person was a nonrespondent for programming ease. This made the run time longer which was not a concern. The jackknife variance estimate is computed using equation (4).

For comparison purposes this simulation is run again using a jackknife variance estimator that treats imputed values the same as observed values (no re-imputation for each jackknife replicate). Once the imputation is done for a sample, treat the imputed values as observed and then apply equation (4) as in simulation 1.

3. Results

Table 1 gives the results for the estimator $\hat{N}$ of the population size of 1000. Column (1) is the average estimate over the 1000 simulated samples. Column (2) is the empirical variance of $\hat{N}$. This is treated as the target value for jackknife variance estimation. Column (3) is the empirical standard error (square root of column (2)). Column (4) is the relative empirical standard error (column (3) divided by column (1)). Note that columns (2), (3), and (4) are the same for re-imputation and no re-imputation since the population estimate does not depend on variance estimation. Column (5) is the average of the 1000 jackknife variance estimates and column (6) is the ratio of this value to the target empirical variance. Column (7) is the square root of column (6).

Table 1 Average and Empirical Variance of $\hat{N}$

	Mean $\hat{N}$	Variance $\hat{N}$	SE $\hat{N}$	$\frac{SE\hat{N}}{Mean\hat{N}}$	Mean VARjk	Ratio Jack/Target	SQRT Ratio
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Simulation 1	1017	9735	99	0.097	11421	1.173	1.083
Simulation 2							
re-impute	1049	17368	132	0.126	30368	1.748	1.322
no re-impute	1049	17368	132	0.126	17115	0.986	0.993
Simulation 3							
re-impute	1012	13394	116	0.115	14958	1.117	1.057
no re-impute	1012	13394	116	0.115	14632	1.094	1.046

For Simulation 1, which assumes no missing data the mean population estimate is about 1017 for the population of 1000 (overestimate of 1.7%). The jackknife standard error estimate overstated the empirical standard error by 8.3% (one tailed P value < .001).

Simulation 2 was run twice, once with re-imputation for each jackknife replicate and once treating the imputed value as observed with no re-imputation for each jackknife estimate. The population is overestimated by about 5%. Simulation 2 imputes for missing data by using the fitted imputation logistic regression from the respondents and a uniform random (1,0) draw to impute either match or nonmatch for each nonrespondent. The bias in population estimation is about 3 times larger than Simulation 1 which assumed no missing data. When re-imputation is done for each jackknife replicate, the jackknife standard error estimate overstates the empirical standard error by 32.2% (one tailed P-value < .001). When the imputed values are treated as observed, the jackknife variance estimate has practically no bias (0.7% underestimate; not significant at 10% level).

Simulation 3 was run twice, once with re-imputation for each jackknife replicate and once treating the imputed value as observed with no re-imputation for each jackknife estimate. The population is overestimated by about 1.2%. Simulation 3 imputes for missing data by using the fitted imputation logistic regression from the respondents to create two records, one for a match and one for a nonmatch for each nonrespondent and uses weighted logistic regression. The bias in population estimation is slightly less than for Simulation 1 which assumed no missing data. When re-imputation is done for each jackknife replicate, the jackknife standard error estimate overstates the empirical standard error by 5.7% (one tailed P-value = .03). When the imputed values are treated as observed, the jackknife variance estimate overstates the empirical standard error by 4.6% (not significant at 10% level).

The major difference between Simulation 2 and Simulation 3 is that Simulation 2 imputes a value (match or nonmatch) using the imputation logistic regression model and a draw from the Uniform (0,1) distribution, while Simulation 3 creates 2 records for each nonrespondent to represent both a match and nonmatch and creates appropriate weights. The population estimate empirical variance is about 30% higher (standard error about 14%) for Simulation 2 due to using a random draw from the Uniform(0,1) distribution. For jackknife variance estimation when a value is imputed (Simulation 2), it appears to be much better to treat the imputed values as observed for jackknife variance estimation. For jackknife variance estimation when two records are created for nonrespondents (Simulation 3), it is slightly better to treat the imputed values as observed for jackknife variance estimation. The jackknife variance estimator that has a re-imputation produces an acceptable error with an overestimate of standard error slightly less than 6%.

4. Summary

For this population with simple random sampling, jackknife variance estimation when there is no missing data overstated the standard error by about 8%. If a random draw was used with logistic regression to impute a status of match or nonmatch for simulated nonrespondents, a simple jackknife treating the imputed values as fixed produced very accurate variance estimation while re-imputing resulted in about a 30% overestimate of standard error. If two records were created to represent both a match and nonmatch with appropriate weights based on an imputation logistic regression model, simple jackknife variance estimation treating the imputed results as observed and including a re-imputation for each jackknife replicate produced similar results(about a 5% overestimate of standard error).

The conclusion from this analysis would clearly be to create two records for nonrespondents to reduce true variance and to treat the imputed values as observed for jackknife variance estimation. However, re-imputing for each jackknife replicate produces similar jackknife variance estimation results.

For the 2010 Census Coverage Measurement (CCM) Survey we are planning to create two records for nonrespondents. An empirical study using real Census 2000 data and the sample design actually used for CCM compared the simple jackknife variance estimate of logistic regression parameters with "true" measures of variance based on Monte Carlo variances. These jackknife variance estimates were reasonable so that the plan will be to treat the imputed values as fixed for jackknife variance estimation for the 2010 CCM.

5. References

Alho J.M. (1990), "Logistic Regression in Capture-Recapture Models", "Biometrics 46, 623-635.

Lee H., Rancourt E., and Sarndal, C. (2002), "Variance Estimation from survey Data under Single Imputation," Survey Nonresponse, John Wiley and Sons, 315-328.